LLNL-BOOK-425971

“The Design of Genomic Signatures for Pathogen Identification and Characterization”

Tom Slezak, Shea Gardner, Jonathan Allen, Elizabeth Vitalis, Crystal Jaing

Livermore National Laboratory, Livermore CA

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Genomic signatures

This chapter will address some of the many issues associated with the identification of *signatures* based on genomic DNA/RNA, which can be used to identify and characterize pathogens for biodefense and microbial forensic goals. For the purposes of this chapter, we define a signature as one or more strings of contiguous genomic DNA or RNA bases that are sufficient to identify a pathogenic *target* of interest at the desired *resolution* and which could be instantiated with particular *detection chemistry* on a particular *platform*. The *target* may be a whole organism, an individual functional mechanism (e.g., a toxin gene), or simply a nucleic acid indicative of the organism. The desired resolution will vary with each program's goals but could easily range from family to genus to species to strain to isolate. The resolution may not be taxonomically based but rather pan-mechanistic in nature: detecting virulence or antibiotic-resistance genes shared by multiple microbes. Entire industries exist around different detection chemistries and instrument platforms for identification of pathogens, and we will only briefly mention a few of the techniques that we have used at Lawrence Livermore National Laboratory (LLNL) to support our biosecurity-related work since 2000. Most nucleic acid based detection chemistries involve the ability to *isolate* and *amplify* the signature target region(s), combined with a technique to *detect* the amplification.

Genomic signature based identification techniques have the advantage of being precise, highly sensitive and relatively fast in comparison to *biochemical* typing methods and *protein* signatures. Classical biochemical typing methods were developed long before knowledge of DNA and resulted in dozens of tests (Gram's stain, differential growth characteristics media, etc.) that could be used to roughly characterize the major known pathogens (of course some are uncultivable). These tests could take many days to complete and precise resolution of species and strains is not always possible. In contrast, protein recognition signatures composed of *antibodies* or *synthetic high-affinity ligands* offer extremely fast results but require a large quantity of the target to be present. False positives/negatives are also a problem with some protein-based techniques (home pregnancy kits use this basic approach).

Different types and resolutions of genomic signatures

Genomic signatures can be intended for many different purposes, and applied at multiple different resolutions. At LLNL, we have been working on signatures that could be broken out into several categories: 1) Organism signatures, 2) Mechanism signatures, and 3) Method signatures.

Organism signatures are intended to uniquely identify the organism(s) involved. *Mechanism signatures* can be best thought of as identifying particular genes that result in functional properties such as virulence, antibiotic resistance, or host range. The primary reason to identify mechanisms, independent of organisms, is to detect potential genetic engineering. A secondary reason is because nature has shared many important mechanisms on its own over the millennia, and thus they may not be sufficiently “unique” to identify specific organisms. Knowledge of whether a particular isolate has the full “virulence kit” or possesses unusual antibiotic resistance properties and whether it is human-transmissible is important for biodefense and public health response. *Method signatures* present yet another dimension of analyzing pathogens: Evidence of potential bacterial *genetic engineering* may be seen in a genome by checking for traces of the bacterial *vector(s)* that may have been used to insert one or more foreign genes and related components (promoters, etc.) into the genome being modified. In the future, *host range signatures* might indicate that an otherwise uncharacterized pathogen was potentially capable of evading or defeating the immune system of a particular host organism.

Potential target organisms

Genetic signatures can be used to identify any living organisms and viruses that contain intact DNA or RNA. Focusing on biosecurity, we are primarily interested in identifying bacteria, viruses, and fungi that could potentially be used to threaten human, animal, or plant life, to disrupt our economy, or to disturb our social order. Note that there is a wide range of genome sizes involved. RNA viruses are generally small (Foot and mouth disease virus is about 8 Kbp, SARS coronavirus is about 30 Kbp), while Variola virus (causative agent of smallpox) is a large DNA virus at about 200 Kbp. High-threat bacterial pathogens tend to be in the 2-5 Mbp size range (*Yersinia pestis*, causative agent of plague, is about 4 Mbp while *Bacillus anthracis* is about 5 Mbp.) Fungi can range from 10 Mbp to over 700 Mbp. As can be imagined, the sequencing databases have many more viral genomes than bacterial, and many more bacterial genomes than fungal. In comparison, the human genome is about 3 Gbp and wheat is about 16 Gbp.

Signature resolution

Organism detection signatures must be *conserved* sequence, reliable and able to detect all intended organisms to minimize false negatives; and *unique* sequence, specific to the target organism and not detecting non-target organisms to minimize false positives. Organism detection signatures can be at different *taxonomic resolution*, typically genus, species, or strain.

In biosecurity applications, high resolution signatures are needed to precisely identify particular isolates or strains. In past years, a large distinction was drawn between *identification* or *detection* signatures and *forensic* signatures, where forensic signatures were typically thought of as at the strain level or below (typically thought of as *sub-strain* or *isolate-specific*). More recently the distinction has become blurred because taxonomic distinctions have become less certain, and because new signature techniques provide increased resolution levels. Using current commercially-available microarray technologies that allow several millions of signatures to be designed on each chip, one can interrogate the entire resolution range (genus, species, strain, and isolate) for desired pathogen targets, providing both detection and forensic resolution. Signature design today is a combination of the desired signature purpose, our current understanding of the diversity of the organism being targeted, and the particular mission constraints that may dictate the detection chemistry and platform to be used for either biodefense or public health.

Genomic sequence data: what to use, and where to get it

There is no single resource for all the genomic sequence data pertinent to signature design. The most comprehensive public source for genomic sequence data is GenBank which is located at the National Center for Biotechnology Information (NCBI) website [**http://www.ncbi.nlm.nih.gov/**]. NCBI has reciprocal data exchange agreements with EMBL in the United Kingdom and DDBJ in Japan, which are equivalent databases used heavily in those parts of the world. Most authors of published sequence data usually submit a final version of their sequence data sets to GenBank. However, numerous sequence databases exist that have organism specific data that may not be found in GenBank during the interim period of data generation and manuscript preparation and those sites would need to be probed directly to obtain the most recent and up-to-date sequence data. Some examples of these publicly available resources are the Integrated Microbial Genomics (IMG) project at the Joint Genome Institute (http://img.jgi.doe.gov), The Comprehensive Microbial Resource (CMR) at the JC Venter Institute (the institute formerly known as TIGR, http://cmr.jcvi.org), and The Sanger Institute in the United Kingdom (http://www.sanger.ac.uk).

Sequence data most useful for signature design falls into two major categories: finished and draft data of *isolated organisms*. Draft genomes are comprised of multiple sets of overlapping reads, called *contigs*, potentially with little or no information about the order or orientation of the contigs relative to the original genome. Draft sequence is often described by a depth factor which is a numeric statement about the average redundancy of coverage at any base position, and thus confidence. A 3X draft sequence would have, on average, at least 3 overlapping reads that contain each base in the genome being sequenced; 8-10X depth is a common stopping point for draft genome data generation for traditional Sanger sequencing (where read lengths averaging 800+bp are common). More recent generations of sequencing based on pyrosequencing technology yield shorter reads (100-200+bp for Roche 454 sequencers and 32-75+bp for the machines from Illumina or Life Technologies), and may feature depths of 50X or greater.

Finished whole-genome microbial sequences have undergone an iterative process to assemble contigs and then use a variety of techniques to order and orient them and close any gaps. This often lengthy and costly process, when completed, produces a single string of high quality bases from the individual and scrambled contigs of the draft sequence. Obviously, finished genomes are superior to drafts when it comes to performing *annotation* of gene content or other features, as well as for performing *multiple sequence alignments* to compare 2 or more genomes. In our experience at LLNL, an 8-10x Sanger draft genome provides sufficient information for DNA signature design purposes [**Gardner et al., NAR 2005**]. When you consider that finished microbial genomes can be

4-10 times as expensive as draft, it is not surprising that many microbial genomes will never be finished. Increasingly, short-read sequences are being *mapped* to reference genomes in lieu of a *de novo assembly*.

Another increasingly important category of data is *metagenomic sequence*, where no attempt has been made to isolate individual organisms for sequencing. Sometimes this is because no way is known to isolate and culture the particular organism(s) of interest. Only a tiny fraction of organisms can be cultured *in vitro* and our knowledge base is greatly skewed towards those that can. At other times it is because what is desired is a sampling of an entire community of organisms. Although numerous metagenomic samples have been sequenced, it is exceedingly rare for complete assemblies of sequence from multiple organisms to result. One exception is a very small symbiotic bacterial community found living in an extremely harsh acidic environment in a mine [**Allen et al., 2007**]. Metagenomic data is not currently of much utility for genomic signature development. A recent paper on the acid-mine bacterial community is providing clues about the evolution of viral resistance [**Banfield, 2008**] which illustrates the vital role metagenomic sequencing will play in expanding our *systems biology* knowledge at both the organism and ecosystem level.

Searching for sequence data based on free-text queries can be problematic. For example, GenBank does not enforce consistency with sequence designation. Not all complete genomes have “complete genome” in the title, and some that do are not actually complete genomes. We have encountered complete genomes that were labeled “complete cds” (coding sequence), “complete gene”, or otherwise unlabeled as a complete genome. Curation is required to validate any sequence data that is obtained from a public resource and periodic “in-house” testing against benchmark data is necessary to maintain a database of high fidelity. A related problem is distinguishing when a new finished genome should replace a prior draft, as strain name, authors, or institutions may have changed.

Identifying conserved sequence among targets

Finding regions of conservation across all target genomes can be done with *alignment-based* methods, and with *alignment-free* methods. The difference between methods revolves around a tradeoff between time and quality.

The first issue to be faced when searching for conservation with a multiple-sequence alignment (MSA) is the amount of sequence (*breadth*) that an alignment method can handle. Alignments sometimes fail when the input sequences are very long or when there are a large number of sequences to be aligned (*depth*), even if the sequences are not particularly long. Failure happens because a MSA takes impractically long to finish, due to the intractable computational complexity involved, or due to a lack of memory. These limitations mean the optimal alignment approach may vary depending on the breadth and depth of sequences used as input. The recent explosion of genome sequence data has resulted in a lack of MSA algorithms that can scale appropriately.

Alignment-free methods for finding consensus can be a shortcut if a complete MSA is impractical or not needed for downstream analysis. Building an alignment-free consensus relies on one sequence serving as a reference for the sequence order of the remaining sequences. This reference sequence is compared pair wise with the remaining sequences, and the consensus is expressed in the sequence order of the reference. This is often less computationally complex than performing a complete MSA, and the results are of sufficient quality to identify suitably conserved regions for potential signatures.

Another topic of concern when identifying conserved sequence regions is whether or not an approach can incorporate incomplete and/or draft sequences. Incomplete sequences do not cover the complete genome of the organism. Draft sequences may or may not cover the complete genome and may be of lower quality, particularly near the ends of contigs. Increasingly, the number of genomes being finished to completion is significantly fewer than the number of genomes that will remain incomplete and in draft form. MUMmer [**Kurtz et al., 2004**] is a notable MSA program in this respect because it can align draft and complete genomes. Note that any use of incomplete genomes carries an inherent risk because regions not present in the incomplete genome(s) will not appear to be conserved and thus may not be considered for signature mining.

Finally, viruses are often highly divergent at the nucleotide level. This extreme divergence, common among many RNA viruses, can cause even alignment-free methods that rely on pair wise sequence search to fail at finding all shared genetic regions. Some non-viral organisms have also been observed with enough divergence to make using alignment-free methods error prone. To help overcome the hurdles of divergent targets, we have developed a novel method of signature generation, *Minimal Set Clustering*, described later.

Identifying sequence unique to targets

Finding regions of sequence unique to the target organism is done by searching large sequence databases. There is a tradeoff in sequence search between execution time and search sensitivity. *Heuristic* algorithms (methods that take reasonable short-cuts which may decrease sensitivity) offer the best time performance. *Non-heuristic* algorithms (methods that guarantee complete coverage within the problem space) can have more sensitive results than heuristics, but are slower, and the additional sensitivity is not always significant.

Heuristics are most commonly used because they make it possible to search extremely large databases such as NCBI’s NT (not non-redundant nucleotide database) quickly. The most popular of these is BLAST [**Altschul et al., 1990**], which can scale to provide fast results with large databases by splitting the search space into many parallel processes across compute clusters. If additional limitations in search sensitivity are acceptable, other approaches such as suffix tree based Vmatch [**http://www.vmatch.de/**] can be faster. Another heuristic approach is to compute Hidden Markov Models that represent the sequence families of interest, like in the program HMMER [**http://hmmer.janelia.org/**].

Mining for signatures

After pathogen target regions that are both conserved and unique are found, they are mined for detection signatures. Signatures are found by searching for oligos with appropriate length, melting temperature, and GC ratio, and by searching for oligo combinations with appropriate overall amplicon size, and minimal inter-oligo hybridization potential. Programs such as Primer3 [**Rozen et al., 2000**] can perform some or all of the signature selection work given a target sequence input. Primer3 can be integrated into any signature development pipeline, unlike other packages that only offer a manual graphic interface.

How KPATH signatures are designed

We will now discuss the major design criteria that the LLNL KPATH [**Slezak et al., 2003**] signature design pipeline was built around. KPATH's native signature format which we will describe is TaqMan® PCR. Its ability to handle several other formats will not be described here.

The process begins by looking across all complete target genomes for sequence conservation. We use an in-house, alignment-free, BLAST-based program for finding conservation (unpublished).

Conserved regions of the target genomes are next screened across our complete genome database in search of potential cross-reactions. Since the oligos of TaqMan® signatures are about 18 bp to 30 bp long, a fairly large seed-length of 18 is acceptable (which means that some short perfectly matching sequences may be omitted from the results). Larger seed-lengths make it possible for us to search much larger databases in reasonable amounts of time. We currently use Vmatch for large database searches.

The resulting conserved and apparently unique sequence, which has no significant similarity to other known sequence, is now mined for signatures. It is important to note that we only find *apparent uniqueness* based on the state of the current whole-genome database available to us. We anticipate that as additional pathogen targets, near-neighbor organisms, and other organisms are sequenced, our regions of conservation and uniqueness will diminish. For this reason, *signature design is an iterative process and not an endpoint*. The original KPATH system used Primer3 in a single execution to identify TaqMan® signature candidates with a forward primer, reverse primer, and a hybridization probe. To let us enforce additional signature design constraints and options without ruling out potential target regions, we converted signature identification into two executions of Primer3 – one for primer pairs, and one for probes. The separate primer and probe results are combined with an in-house signature builder and scorer to allow us to identify the best combinations of primers and probes.

Next, signatures are filtered down so there is little or no overlap of candidate signatures within the target organism. When exhaustive signature searches are performed, many of the mathematically best signature candidates will share oligos and generally be very similar. This means that choosing the best scoring signatures for any given locus helps us remove excess redundancy from the pool of signature candidates.

The final check we typically perform is a TaqSim [**http://staff.vbi.vt.edu/dyermd/publications/taqsim.html**] comparison of all signature candidates against NCBI's NT database. This highly-sensitive BLAST search TaqMan® PCR simulator with post-processing lets us verify that the signature candidates are conserved enough to detect all the expected targets, and unique so that there are no non-target hits. Depending on the intended uses of the signatures (e.g., environmental versus clinical samples) we may choose to do additional testing against genomes from human or other complex organisms.

We note that in recent years other DNA signature pipelines have been built that take a reverse approach. Like LLNL's Minimal Set Clustering described, they first generate all potential valid TaqMan® PCR signatures for each available genome of a target organism and then BLAST them to check for sufficient conservation and uniqueness.

RNA viruses present additional challenges

High rates of mutation and lack of genome repair mechanisms in many viruses generate high levels of intraspecific diversity and result in *quasispecies*, particularly for many single-stranded RNA viruses. Consequently, PCR-based signatures for viral detection often require high levels of degeneracy or multiplexing in order to robustly detect all variants. Large amounts of sequence data are often required to represent the range of target diversity, sometimes dozens to hundreds of genomes. As noted previously, building multiple sequence alignments with many diverse genomes taxes the capabilities of most available software. Once an alignment is built, it may reveal insufficient consensus for even a single primer, much less a pair, to detect all members of some species (e.g.: Human immunodeficiency virus-1 or Influenza A).

One solution is to subdivide the targets into smaller or more closely related subgroups such as *clade*, *serotype*, or *phenotype* of interest (examples of phenotypes could include: virulent versus vaccine, domestic versus foreign), and attempt to find signatures separately for each subgroup. This approach implies that multiple signatures will be required for species level detection of all subgroups. One must make an assessment in advance of signature design of how best to subdivide the target sequences. A second approach is to allow *degenerate* or *inosine* bases, so that a single signature will detect more diverse genomes. Specificity may suffer if some combinations of degenerate bases also pick up non-target species. Sensitivity may decline since the specific priming sequence for a given target is diluted in the degenerate mix. A number of tools which require a multiple sequence alignment as input are available for degenerate primer design (e.g. SCPrimer, PrimaClade, Primo, Amplicon, and HYDEN). A third approach is to forego sequence alignment altogether, and to look for sets of primer-length oligomers of length k, or *k-mers*, present in many targets and unique relative to non-target sequences. Using combinatoric or greedy algorithms, one can build a signature set of k-mers such that each target contains at least two k-mers to function as forward and reverse primers. This approach demands large amounts of computing memory to store all candidate k-mers for large or many genomes, especially as k increases above 20, and may require suffix trees or other techniques for data compression.

A fourth approach that we have employed is called Minimal Set Clustering (MSC). It avoids the need for a multiple sequence alignment or a priori sub grouping of the target sequences, so this method can be run “blindly” without expert knowledge of the target species. It begins by removing non-unique regions from consideration as primers or probes from each of the target sequences relative to a database of non-target sequences. The remaining unique regions of each target sequence are mined for all or many candidate signatures, without regard for conservation among other targets, yet satisfying user specifications for primer and probe length, T_m, GC%, amplicon length, etc. All candidate signatures are compared to all targets and clustered by the subset of targets they are predicted to detect. To predict detection, we may require that a signature’s primers and probe have a perfect match to target in the correct orientation and proximity, or we may relax the match requirements to allow a limited number of mismatches, so long as T_m remains above a specified threshold or those mismatches do not occur too close to a primer’s 3’ end. Signatures within a given cluster are equivalent, in that they are predicted to detect the same subset of targets, so by clustering we reduce the redundancy and size of the problem to finding a small set of signatures that detect all targets. Nevertheless, finding the optimal solution of the fewest clusters to detect all targets is an *NP complete* problem, so for large data sets we use a greedy algorithm to find a small number of clusters that together should pick up all targets. LLNL has used this method to design signature sets for numerous RNA viruses, including Influenza A HA serotypes, Foot and Mouth disease, Norwalk, Crimean-Congo hemorrhagic fever, Ebola, and other divergent viruses. The figure below shows the result of an MSC computation for Crimean-Congo hemorrhagic fever performed in 2005, with the resulting signatures displayed against a whole-genome phylogenetic tree of all the sequences available at that time.

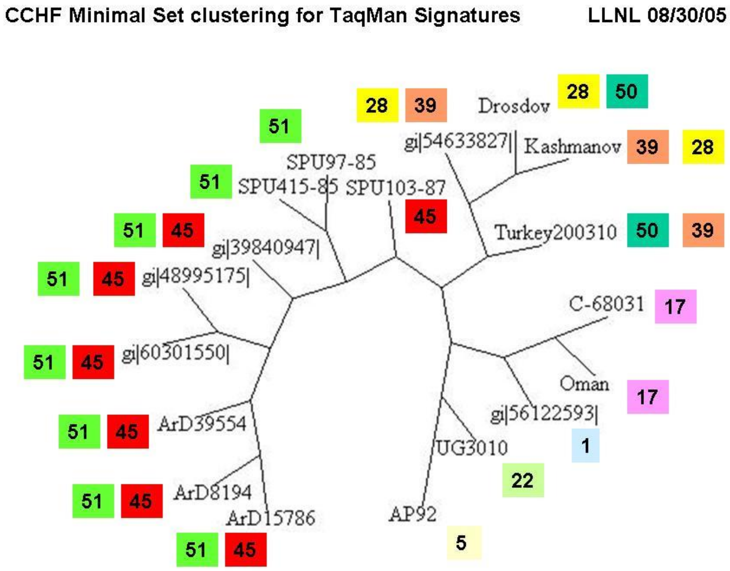

Figure 1. The result of Minimal Subset Clustering signatures for Crimean-Congo Hemorrhagic Fever virus (CCHFV) displayed against a whole-genome phylogenetic tree of the available target genomes. Note that signatures 45 and 51 cover a wide range of isolates from one geographical location, while signatures 28, 39, and 50 cover isolates found in Eastern Europe. Signatures 1, 5, 17, and 22 are required to detect some historical isolates that are not likely to be in current circulation.

Signatures of potential bacterial genetic engineering

Detecting evidence for genetic engineering in bacteria is challenging when the target modification is not known and the effects of an outbreak on human health are not well understood. We may, for example, anticipate a biological outbreak that employs a bacterial host containing a foreign toxin, but the observed effects of the toxin may not implicate a known gene. Even in cases where the gene is known, it may be difficult to rule out a natural origin for the outbreak. In such cases, it may be useful to search for more direct evidence of the genetic engineering tools used to insert and express foreign genes in a bacterial host. Among the most widely used and readily available tools for genetic engineering in bacteria are the *artificial vector* DNA molecules.

Genetic engineering with artificial vectors began with efforts to improve on the early work using *natural plasmids* for gene cloning. Natural plasmids are extra chromosomal replicons (self replicating molecules) which come in both circular and linear form and are generally non essential genetic material for the bacterial host but can confer important phenotypes such as virulence and drug resistance. These plasmids are *mobile genetic elements* that serve as a natural mechanism for exchange of genetic material across different bacterial species [**Frost et al., 2005**]. The artificial vectors are natural plasmid derivatives designed to improve support for the insertion and manipulation of foreign genetic elements in the carrier plasmid.

We use the term “artificial vector” to refer to replicons created through human intervention to explicitly distinguish them from their natural plasmid precursors. The sequence features designed to support genetic manipulation form the basis for the methods used to distinguish artificial vector sequence from natural plasmids. The most common artificial vector specific feature is the *multiple cloning site region*, which is a sequence insert containing clusters of restriction enzyme sites used to facilitate insertion of the foreign gene elements. *Selection marker genes* also play an important role in selecting bacteria, which maintain the artificial vector. The *gene transcription control unit*, which includes a promoter sequence and transcription terminator sequence for the foreign gene elements are also important features along with the *origin of replication site* required for maintenance of the artificial vector in the bacterial colony [**Solar et al., 1998**].

Detecting artificial vector sequence in a mixed bacterial sample potentially requires testing a broad range of sequence targets. This suggests the use of an assay with a high degree of multiplex capability that tests for the presence of a large number of sequences simultaneously. Microarray based assays are a logical choice for accommodating a large number of artificial vector detection probes. The large collection of artificial vector sequences can be clustered according to exact *k-mer* sequence matching to find the k-mers shared among different vector sequence [**Allen et al., 2008**]. The sequence length k corresponds to the desired probe length used in the microarray design. Each cluster of shared sequence is compared against all available sequenced natural chromosomal bacterial and viral genomes including natural plasmids to identify which k-mers in the artificial vector sequence are distinct from the natural background. These unique k-mers are called *candidate signatures*. After candidate signatures are found, a probe set is created that ensures that each vector contributes a preset minimum number of candidate signatures to the final microarray probe set design. A greedy algorithm can be used to pick the signatures shared by the greatest number of artificial vectors, selecting candidate signatures in decreasing order.

Additional post processing steps may further improve the quality of the signature probe set design to achieve the ultimate goal of sensitive detection, while maintaining a hybridization pattern on the microarray that distinguishes the artificial vectors from the natural background found in a mixed sample. Once the initial probe set is designed, a BLAST search can be used to tune the probe set by replacing the candidate signatures with near matches to the background with candidates showing a greater percentage of vector unique variation. Cross-validation can be used to estimate a similarity threshold for distinguishing the artificial and natural genomic sets. (An example of this approach using cross-validation is given in **Allen et al., 2008**.) Another post processing step is to tune the probe set to ensure probes derived from each vector come from multiple functional regions. Confidence in vector detection is boosted when probes are found for multiple functional locations. Using probes from multiple regions may also provide useful forensic information on the origins and function of the detected artificial vector.

Given the similarities between artificial vectors and natural plasmids, having additional probes for natural plasmids allows for a direct comparison with the natural plasmid hybridization pattern, which could reduce the potential for false positive predictions.

Viral and Bacterial Detection Array

Numerous microarrays have been designed for viral discovery, detection, and re-sequencing [**Wang et al., 2002**], [**Palacios et al., 2007**], [**Lin et al., 2006**], [**Jabado et al., 2008**]. Re-sequencing arrays can provide sequence information for viruses closely related (>90% similarity) to the sequences from which the array was designed. Discovery arrays to detect more diverse and more distantly related organisms have been built using techniques for selecting probes from regions of known conservation based on BLAST nucleotide sequence similarity [**Wang et al., 2003**] or profile HMM and motif indications of amino acid sequence conservation [**Jabado et al., 2008**]. Array design to span an entire kingdom on a single microarray demands substantial investment in probe selection algorithms. LLNL designed a microarray to detect all bacteria, plasmids, and viruses, based on all available whole genome, whole segment, and whole plasmid sequences, and are in the process of including probes for highly conserved fungal genes as well. We attempted to find probes that are unique to each viral and bacterial family, and favor probes conserved within a family. We used probes 50-65 bases long, enabling sensitive detection of targets with some sequence variation relative to the probe. We used a greedy minimal set cover algorithm to ensure that all sequences have at least 50 (for viruses) or 15 (for bacteria and plasmids) probes per sequence. We allowed some mismatches between probe and target, based on previous mismatch experiments in which we determined that probes with a contiguous match at least 29 bases long and with 85% sequence similarity between probe and target still gave a strong signal intensity. Our design should characterize unknowns to at least the family level, and in all cases tested so far, including blinded clinical samples containing multiple viruses, we are able to accurately detect and characterize all viruses contained in that sample to the species or strain level [**Gardner et al., 2010**].

Our first generation viral array included 36,000 probes designed from family-specific regions of all 72 viral families, and our second version included 170,000 viral probes, again from family-specific regions. There were no regions greater than 25 bp matches to human or bacteria and no regions greater than 17 bp matches to other non-target viral families. In addition, we also included the 20,000 probes from the Virochip developed by Dr. Joseph DeRisi from UCSF as a control [**Wang et al., 2002**].

Our preliminary testing using NimbleGen arrays with mixed DNA and RNA viruses and with blinded clinical samples showed accurate detection of multiple viruses in a single sample. In addition, we can identify the exact strains and isolates hybridized as a mixed sample, though the array was designed to guarantee discrimination only to family. We developed a novel statistical method that is based on likelihood maximization within a Bayesian network, incorporating a sophisticated probabilistic model of probe-target hybridization developed and validated with experimental data from hundreds of thousands of probe intensity measurements. The method is designed to enable quantifiable predictions of likelihood for the presence of each of multiple organisms in a complex, mixed sample, which is especially important in an environmental sample or one with chimeric organisms. Our future detection chip designs will include probes from conserved regions of bacterial families and plasmids and fungal families. This chip will be a major platform for identification of known and unknown pathogens.

The future of genomic signatures

Issues related to scaling, taxonomy, and technology advances appear to be the main drivers for the future of genomic signatures.

Scaling problems all stem from the exponential rate at which genomic sequence data is growing. Although it is inexpensive to buy sufficient hardware to physically store the data, the current generation of bioinformatics tools was designed in an era when it was a luxury to have a handful of genomes of a particular pathogen available to work with. In recent years the Influenza Community Sequencing Project [**Ghedin et al., 2005**] has deposited many thousands of complete influenza genomes into GenBank, far exceeding the capacity of most tools to handle them. Similarly, some of the new sequencing technologies can generate billions of bases in a single run from metagenomic samples [**Mardis, 2007**], but truly efficient software that takes full advantage of this information is lacking. It will likely take several years for research funding to be properly focused to close this bioinformatics tool gap. Another aspect of scaling problems is that few researchers have access to computers with large enough memories to be able to process certain classes of sequence analyses related to genomic signature design. Computer clusters that are optimal for physical science problems (where each node represents a point in a 3-D physical grid representation and almost all communication is with the nearest-neighbor nodes) are sub-optimal for some classes of biological sequence algorithms where a large memory computer would be better.

Earlier we mentioned difficulties with the evolving taxonomy of pathogenic organisms as classification schemes originally developed based on phenomenology are faced now with genomic inconsistencies. The current flood of metagenomic data is presenting us with an even larger problem: what exactly do concepts like “species” and “strains” mean if it turns out that microbial life is a broad spectrum with few well-defined transitions? It is now common to refer to a “core genome” and additional distinct gene content variation that presumably is responsible for different phenotypes [**Nelson, 2004**]. It is possible that new concepts and terminology will be needed to map existing taxonomic categories into the genomic reality of the 21st century.

The rate of advancement in sequencing technology exceeds that even of computers, fueled by the promise of *personalized medicine* if individual drug and disease reactions can be determined, and if individual genetic variation can be efficiently determined via low-cost sequencing. The field of pathogen diagnostics is riding this technology wave, too small a market to have any direct influence. Note that the read lengths of some new sequencing technologies may be too short to provide confident pathogen identification based on a single read, meaning that direct metagenomic identification of human pathogens from complex clinical or environmental samples contains some degree of uncertainty. Microarrays will have to ride their own faster/cheaper/more-information-per-chip curve if they are not to become obsolete within a few years. Alternatively, one could argue that future advances in protein detection technology could lead to breakthroughs in fast "dipstick" assays (similar to current home pregnancy test kits) that could provide fast, accurate, and inexpensive results for pathogen detection. In all likelihood, all these techniques will continue to compete as they evolve asynchronously.

Another technological advance is seen in the recent breakthroughs in gene and genome synthesis [**Gibson 2008**]. Not only do we need to deal with emerging natural viruses from every remote corner of the planet, but now we also need to deal with the fact that for relatively modest amounts of money, it is possible to synthesize combinatorial versions of any DNA one might wish to (re)create. This potential ability to create a new class of supercharged pathogens, as well as the possibility of synthesized pathogens that do not exist in nature, puts a new urgency into ensuring that we have adequate tools to deal with these evolving biothreats.

What all this means for genomic signature design is that we will have to exist in a combination of a data avalanche, new analysis tools, and rapidly evolving new technologies. Against this background of change, we will have to deal with new missions and new challenges from adversaries equipped with the latest technologies. Fittingly for biodefense, it is indeed a very Darwinian challenge that faces us.

REFERENCES

Allen, E.E., Tyson, G.W., Whitaker, R.J., Detter, J.C., Richardson, P.M., and Banfield, J.F. (2007) Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. USA* **104**(6): 1883-1888.

Allen, J.E., Gardner, S.N., and Slezak, T.R. (2008) DNA signatures for detecting genetic engineering in bacteria. *Genome Biology* **9**(3): R56.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**: 403-410.

Banfield, J.F., and Andersson, A. (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **230**: 1047-1050.

Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005) Mobile Genetic Elements: The Agents of Open Source Evolution. *Nature Reviews Microbiology* **3**: 722-732.

Gardner, S.N., Lam, M.W., Smith, J.R., Torres, C.L., and Slezak, T.R. (2005) Draft versus finished sequence data for DNA and protein diagnostic signature development. *Nucleic Acids Research* **33**(18): 5838-5850.

Gardner, S.N., Jaing, C., McLoughlin, K., et al. (2010) A Microbial Detection Array (MDA) for Viral and Bacterial Detection. *(submitted).*

Ghedin, E., Sengamalay, N.A., Shumway, M., et al. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437**: 1162-1166.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., et al. (2008) Complete Chemical Synthesis, Assembly, and Cloning of a *Mycoplasma genitalium* Genome. *Science* **319**: 1215-1220.

Jabado, O.J., Liu, Y., Conlan, S., et al. (2008) Comprehensive viral oligonucleotide probe design using conserved protein regions. *Nucleic Acids Research* **36**(1): e3.

Kurtz, S., Phillippy, A., Delcher, A.L., et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology* **5**(2): R12.

Lin, B., Wang, Z., Vora, G.J., et al. (2006) Broad-Spectrum respiratory tract pathogen identification using re-sequencing DNA microarrays. *Genome Research* 10.1101/gr.4337206.

Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**(3).

Nelson, K.E., Fouts, D.E., Mongodin, E.F., et al. (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. *Nucleic Acids Research* **32**(8): 2386-2395.

Palacios, G., Quan, P.L., Jabado, O.J., et al. (2007) Panmicrobial Oligonucleotide Array for Diagnosis of Infectious Diseases. *Emerging Infectious Diseases* **13**(1): 73-81.

Rozen, S., and Skaletsky, H. (2000) Primer3 on the WWW for General Users and for Biologist Programmers. *Methods in Molecular Biology* **132**: 365-386.

Slezak. T.R., Kuczmarski, T., Ott, L., et al. (2003) Comparative genomics tools applied to bioterrorism defense. *Brief Bioinformatics* **4**: 133-149.

Solar, G.D., Giraldo, R., Ruiz-Echevarria, M.J., Espinosa, M., and Diaz-Orejas, R. (1998) Replication and Control of Circular Bacterial Plasmids. *Microbiology and Molecular Biology Reviews* **62**(2): 434-464.

Wang, D., Coscoy, L., Zylberberg, M., et al. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. USA* **99**(24): 15687-15692.

Wang, D., Urisman, A., Liu, Y-T., Springer, M., Ksiazek, T., et al. (2003) Viral Discovery and Sequence Recovery Using DNA Microarrays. PLoS Biol 1(2): e2. doi:10.1371/journal.pbio.0000002